DOCUMENT RESUME

ED 372 078 TM 021 709

AUTHOR Spray, Judith A.; Reckase, Mark D.
TITLE The Selection of Test Items for Decision Making with a Computer Adaptive Test.
PUB DATE Apr 94
NOTE 14p.; Paper presented at the Annual Meeting of the National Council on Measurement in Education (New Orleans, LA, April 5-7, 1994).
PUB TYPE Reports - Evaluative/Feasibility (142) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Ability; *Adaptive Testing; Bayesian Statistics; *Computer Assisted Testing; *Decision Making; Estimation (Mathematics); Higher Education; Item Banks; Mathematics Tests; Research Methodology; *Selection; Simulation; Student Placement; Test Construction; *Test Items; Test Length
IDENTIFIERS COMPASS (Computer Assisted Test)

ABSTRACT

The issue of test-item selection in support of decision making in adaptive testing is considered. The number of items needed to make a decision is compared for two approaches: selecting items from an item pool that are most informative at the decision point or selecting items that are most informative at the examinee's ability level. The first approach is explored using a method based on sequential Bayes estimation and the second is based on the sequential probability ratio test (SPRT). Simulation studies used item parameters from an actual mathematics item pool developed for COMPASS, the American College Testing Program's computerized adaptive-testing package for college placement. Results could not be compared directly because matching the decision error rates for the two procedures is beyond the scope of the paper. Results do provide clear evidence that selecting items to maximize information at the decision points results in shorter average test lengths than selecting items to maximize information either at true ability (which is impossible) or at the most recent ability estimate. It is expected that these results will generalize to many adaptive-testing applications. (Contains 14 references and 2 figures.) (SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
☑ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
P.A. FARRANT
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

ED 372 078

The Selection of Test Items for Decision Making with a Computer Adaptive Test[1]

Judith A. Spray
Mark D. Reckase
American College Testing

[1]Paper presented at the national meeting of the National Council on Measurement in Education, New Orleans, April 1994.

TM021705

BEST COPY AVAILABLE

The Selection of Test Items for Decision Making with a Computer Adaptive Test

Over the last few years, applications of adaptive testing have become increasingly more common. The Differential Aptitude Test (DAT) has had an adaptive version for a number of years (Henly, Klebe & McBride, 1989), an adaptive version of the Armed Services Aptitude Battery (ASVAB) is near implementation (Divgi, 1991), both ACT (American College Testing Program, 1993) and College Board (Ward, 1988) have made adaptive tests for placement available, and a number of certification/licensure organizations are considering adaptive tests for their applications (e.g., Bergstrom, Lunz & Gershon, 1992).

With the increasing, and often high stakes uses of adaptive testing, the details of the procedures for implementing adaptive tests are receiving more attention. For example, research is being done on whether examinees should be allowed to change their answers to questions (Lunz, Bergstrom & Wright, 1992) and on the use of response latency information collected during an adaptive test (Parshall, Mittelholz & Miller, 1994).

Among the details of the use of adaptive tests that are being considered is how best to use adaptive tests for making pass/fall decisions. A corollary problem is how to select items so that the decision-making process will be performed efficiently. The issue of item selection in support of decision making is the focus of this paper.

The Problem

Two approaches have been suggested in the literature for selecting test items to yield efficient decisions. The first approach is to select the items from the item pool that are most informative at the decision point. Using this procedure, all examinees start with the same items (unless some sort of exposure control is used) and the only adaptation that takes place is in the length of the test. The test terminates as soon as a decision is made with the desired level of accuracy.

The second approach is to select the items that are most informative at the examinee's most recent ability estimate. This is the procedure that is typically followed when adaptive tests are used to get good estimates of the examinees' ability. The argument for its use is that if accurate ability estimates are obtained, accurate decisions can be made. This procedure is a surrogate for selecting items at the examinee's true ability, but, of course, the true ability is never known.

The purpose of the research reported in this paper is to compare the number of items needed to make a decision using each of the two approaches described above. Although this

seems like a fairly straight forward problem, it is not because the number of items needed to make a decision is related to the decision making procedure that is used. Two different procedures have been suggested in the literature. One is a procedure based on sequential Bayes estimation as implemented by Owen (1975). This procedure has been suggested by Weiss and Kingsbury (1984).

The second procedure is based on the sequential probability ratio test (SPRT) suggested by Wald (1947). Reckase (1983) has suggested the use of this procedure with adaptive testing. Because of the interaction of the decision procedures with the item selection methodology, both decision procedures will be used in this study. However, the results will not be compared directly because matching the decision error rates for the two procedures is beyond the scope of this paper. Within a procedure, the error rates are held constant, even when the number of items administered is different. For a detailed discussion of these issues and a comparison of the classification accuracy of the two procedures, see Spray and Reckase (1993).

A Logical Analysis

Sequential Bayes Procedure

The sequential Bayes procedure, as implemented by Kingsbury and Weiss (1983), selects items for administration that are most informative at the most recent estimate of ability as computed using Owen's (1975) procedure. After each item, a high density interval is estimated assuming a normal posterior distribution using the estimated ability estimate as the mean and the posterior variance as the estimate of dispersion. If the $(1-\alpha)$-high density interval does not cover the decision point, the examinee is classified as either passing or failing, and the test session terminates. If the high density interval includes the decision point, another item is administered, and the posterior distribution is updated. The decision-making process is then repeated.

As described above, the procedure requires that the next item to be selected be the one that is most informative at the most recent ability estimate. But, how would the performance of the procedure change if the items were selected to be most informative at the decision point instead of at the most recent ability estimate? A simple example using the Rasch model provides a rationale for why the two item-selection procedures might not give the same result.

Suppose that the decision point is at $\theta_c = 0.0$ and that the true ability of an examinee is $\theta_t = 1.0$. If the null hypothesis, H_0, is that the examinee is at θ_c, an equivalent null hypothesis is that the proportion of items with $\underline{b} = 0.0$ that the examinee will answer correctly is .5. Thus the decision-making procedure could be to give the examinee items with $\underline{b} = 0.0$, compute the proportion correct, and test the obtained value against .5 using the test statistic

$$z = \frac{.5 - p}{\sqrt{\frac{pq}{n}}}$$

where z is the normal approximation to a binomial,
p is the estimated proportion correct,
q is 1 - p,
and n is the number of items that have been administered.

The observed probability of correct responses for these items assuming the Rasch model holds is expected to be .73 at the examinee's true ability level. If the null hypothesis is tested using $\alpha = .05$, the value of z needs to be less than -1.96 to be significant. To determine the number of items needed to reject the null hypothesis, the following equation can be solved for n.

$$-1.96 = \frac{.5 - .73}{\sqrt{\frac{.73 \times .27}{n}}}$$

The solution is 3.78. Thus, it should be possible to reject the null hypothesis and classify the examinee after about four test items.

If the items are selected at the examinee's true ability of $\theta = 1.0$, the null hypothesis becomes H_0: p = .27, the expected probability of a correct response for persons at the decision point of $\theta = 0.0$ for items with b = 1.0. Of course, the probability of correct response for the examinee will be .5 because the items have been selected to match his/her ability. To determine the expected sample size needed to make the decision in this case, the following equation needs to be solved for n.

$$-1.96 = \frac{.27 - .5}{\sqrt{\frac{.5 \times .5}{n}}}$$

The value of n obtained from the solution of this equation is 4.26. Thus, in this simple case, the decision would be made about a half an item sooner by selecting the items to be maximally informative at the decision point rather than at the examinee's ability level.

If actual decision making with adaptive testing were as simple as this example, no further analyses would be needed. It could be concluded that selecting items at the decision point is more efficient than selecting them at the examinee's level of ability. Of course, reality is much more complicated than this simple example. The examinee's ability is unknown, so items can

not be selected to have maximum information at that point on the ability scale. Instead, some number of items is needed to determine the approximate location on the scale of the examinee's ability estimate.

Further, item pools are not ideal. They have gaps and more items tend to assess middle ranges of abilities than the extremes. In addition, the more complex and nonsymmetric, three-parameter logistic model is often used instead of the Rasch model. The presence of a lower asymptote parameter vastly complicates the analysis. Finally, the sequential Bayes procedure does not lend itself to simple analyses. To account for all of these factors, a fairly detailed simulation study is needed.

Sequential Probability Ratio Test

A logical analysis of item selection for the sequential probability ratio test (SPRT) is much more straightforward. This test is based on a test statistic that is the ratio of the probability of the observed item response vector computed using points on either side of the decision point. If the decision point is θ_c, a region around the decision point is selected such that the examiner is indifferent to the decision that is made for examinees whose true abilities are in the region. In other words, the examinees are so close to the decision point, that their classification can be determined with a coin toss. The bounds of the indifference region are given by θ_0 and θ_1, with θ_0 less than θ_1. The test statistic is

$$L(\mathbf{U}) = \frac{P(\mathbf{U}|\theta_1)}{P(\mathbf{U}|\theta_0)} ,$$

where $L(\mathbf{U})$ is the likelihood ratio for response vector $\mathbf{U}$,
and $P(\mathbf{U}|\theta)$ is the probability of the response vector conditional on the appropriate value of θ.

If the likelihood ratio is sufficiently large, the decision that the examinee is above the decision point is supported. If it the ratio is sufficiently small, the decision that the examinee is below the decision point is supported. The critical values for determining the decision are based on the amount of error in classification that is acceptable.

From the equation, it is clear that decisions will be made quickly when the difference in probabilities for θ_0 and θ_1 is great. That is, when the ICC for the items is very steep between the bounds of the indifference region. This is exactly the case when items are selected to be maximally discriminating at the decision point, assuming the region is not too wide and is symmetric around the decision point, which is usually the case.

As with the sequential Bayes procedure, reality is not as simple as this analysis would suggest. The characteristics of the items pool interact with the functioning of the procedure. The

items that have b-values that are quite different than the scale value of the decision point may be of better quality than the items at the decision point and there may be gaps in the item pool. Use of the three-parameter logistic model also introduces some nonsymmetry to the functioning of the procedure because of the nonzero lower asymptote. However, this analysis suggests that selecting items to have maximum information at the decision point will be more efficient than selecting them at the examinee's ability. A realistic simulation is used to determine whether reality is likely to match the expectations provided by the logical analysis.

The Simulation Studies

Item Pool

The simulation studies used the item parameters from an actual mathematics item pool that was developed for COMPASS, ACT's computerized adaptive testing package for placing students in college courses (ACT, 1993). This item pool had been calibrated using BILOG (Mislevy & Bock, 1984). It contains 200 items that are well spread over the θ-range from -2.0 to +2.0. The means of the item parameters are 1.18, .48, and .16 for the a-, b-, and c-parameters, respectively. The item parameter estimates were treated as known parameters for the purposes of the simulations.

Sequential Bayes Procedure

Three item selection methods were simulated for the sequential Bayes procedure: (1) maximize information at the decision point; (2) maximize information at the examinee's true θ; and (3) maximize information at the examinee's most recent estimate of θ. The last condition was included to determine if more items were needed to make a decision when the examinee's location on the scale first had to be found before the decision could be made. In all cases, ability was estimated using the procedure described by Owen (1975) and testing was terminated when the 95% high density region no longer included the decision point. All testing began assuming a standard normal prior distribution, N(0,1), of ability.

To determine the generalizability of the results to irregularities in the item pool, three decision points were simulated: -.5, 0.0, and 1.0. Examinees were simulated assuming true abilities at .25 intervals from -3.0 to +3.0. The average number of items required to make a decision was computed based on 1000 replications at each θ-level. The maximum test length was set at 50, because it was expected that real applications of adaptive tests were unlikely to use more than 50 items.

Results and Discussion. The results for the three decision points and the three item selection procedures are shown in Figure 1. From the graphs, it can be seen that selecting items to maximize information at the decision point, δ, shown by the solid line, results in the lowest

average test length (ATL) everywhere above a θ-level of about −1.4. Selecting to maximize information at the examinee's true ability, the dotted line, increased the test length slightly, generally less than two items, except for very low ability examinees. The average test length when items were selected to maximize information at the examinee's most recent ability estimate is uniformly greatest, except for low ability examinees. For the decision point of 1.0, the number of items required by this method was much greater, over seven items on average at $\theta = 1.5$.

Several of the results merit further discussion. First, in the range from -1.0 to +1.0, where most examinees will likely be, selecting items to maximize information at the decision point is slightly more efficient than any of the other two options. For this item pool, the prior selected, and the decision points used, the differences are relatively minor in this region, and a user might be willing to accept having to give a few extra items if there were good reasons to select items to maximize information at the most recent ability estimate.

The results for the high ability examinees with the decision point at 1.0 occur because the Bayes estimation procedure must first overcome the N(0,1) prior before a decision can be made. The shift in ability estimates after a correct response is partially dependent on the difference between the prior mean and the item's b-parameter. When items are selected to match the decision point of 1.0 and the prior distribution of ability is N(0,1), there is a relatively large difference between the b-value of the selected item and the initial ability estimates early in the test. This results in quicker movement of the ability estimates to above the decision point than if items are selected at the most recent ability estimate. The ability estimate must move above the decision point before a correct decision can be made.

The lower number of items needed to make decisions for examinees in the -2.0 to -3.0 range is due to the effects of the lower asymptote on the functioning of the procedures. For very low ability examinees, selecting items to match the decision point results in the administration of items that are very difficult for the examinee. For those items, the noise induced by the c-parameter has a prominent effect. If the items are selected to match the current ability estimate, easier items are selected, and the c-parameter has less of an effect.

One final point. Note that when the decision point is at 1.0, many low ability examinees are classified after little more than one test item. This is a result of selecting the prior to be N(0,1). One incorrect response shifts the posterior mean to the left and reduces the posterior variance. Given the prior and the location of the decision point, that one response is sufficient to produce a high density region that does not include the decision point. This may result in high rates of misclassification if high ability examinees make a key entry error.

SPRT Procedure

Two item selection methods were simulated for the SPRT procedure: (1) selecting items to have maximum information at the decision point; and (2) selecting items to have maximum

information at the examinee's true ability. Selecting items to have maximum information at the examinee's most recent ability estimate was not used in this case because ability estimates are not used with the SPRT procedure, and using estimated ability was expected to result in longer average test lengths than using true ability.

The indifference region in all cases was set to be symmetric around the decision point, δ: $\delta \pm .5$. Both the α and β error rates were set at .05. As with the sequential Bayes procedure, 1000 replications were performed at θ-values between -3.0 and +3.0 in increments of .25. The maximum test length was set at 50 items.

Results and Discussion. The results for the SPRT procedure for the three decision points and the two item selection methods are presented in Figure 2. Except for the high ability examinees, the results are similar to those for the sequential Bayes procedure. Generally, selecting items to have maximum information at the examinee's true ability results in longer average test lengths (ATL). This result is quite dramatic for the lower two decision points and examinees above θ of .5. Below $\theta = .5$, the difference in the average test length is generally two or less, with selecting items to match the decision point requiring fewer items to make a decision.

The large difference in number of items for the high ability examinees is a result of the nonzero lower asymptote for the three-parameter logistic model. For example, if an examinee's true ability is 2.0 and the decision point is -.5, the probability ratio that is used as the test statistic is given by

$$L(\boldsymbol{u}) = \frac{P(\boldsymbol{u}|\theta = 0.0)}{P(\boldsymbol{u}|\theta = -1.0)},$$

because the indifference region is from -1.0 to 0.0 (i.e., $-.5 \pm .5$). At the bounds of the indifference region, an item with a b-parameter of about 2.0 will likely have both probabilities near the value of the c-parameter, say .16 and .18. The resulting ratio is very close to 1.0. Thus, many items have to be administered before a decision can be made.

For low ability examinees, this is not the case. For items selected to match their true ability, examinees will likely get about half correct and the other half incorrect. The incorrect responses will have very low probabilities at the bounds of the indifference region, values of -1.0 and 0.0. If they are .02 and .01, respectively for θ equal to -1.0 and 0.0, for example, the likelihood ratio is still .5, a value sufficiently different from 1.0 to assist in making the decision.

Sequential Bayes versus SPRT

The results for the two procedures are fairly consistent. Both show that for examinees in the usually observed range of ability, selecting items to maximize information at the decision

point results in a shorter average test length than the other alternatives considered. When the decision points are high relative to the prior distribution, or for high ability examinees, the differences in average test length for the alternative item selection methods can be quite dramatic. This finding may be quite important when certification or licensure standards are fairly high, or selection criteria are rigorous. Under those circumstances, it is clear that selecting the items to match the decision is the more efficient method.

The average test lengths for the SPRT based and the sequential Bayes procedure are not directly comparable because the procedures have not been matched on error rates. The 95% high density region does not yield the same error rate in making decisions as an α of .05 for the SPRT. The latter error rate is for the entire testing process and applies to classifying above the decision point when an examinee is at the lower bound of the indifference region. The sequential Bayes procedure is not consistent with a hypothesis testing philosophy, and the 95% region is used after each item. Also, the significance tests performed after each item are not independent, further complicating the estimation of the actual error rate. Spray and Reckase (1993) have developed procedures for matching the error rates for the procedures. Under matched conditions, the SPRT is more efficient than the sequential Bayes procedure.

Conclusions

The results of this study provide very clear evidence that selecting items to maximize information at the decision point results in shorter average test lengths than selecting items to maximize information either at true ability (which is impossible) or the most recent ability estimate. The exception seems to be for very low ability examinees when a model with nonzero lower asymptotes is used. For that range of ability, less than -1.5 on the θ-scale, selecting at the estimated ability may save one or two items.

It is expected that these results will generalize to many applications of adaptive testing. The item pool used was not unusually high in quality. The item parameters were typical of those found for commercially prepared tests. The decision points covered a reasonable range, and the results were consistent. The error rates used in the study were similar to those encountered in practice.

Practitioners should be cautious, however, in generalizing these results to applications that use exposure control procedures and complex content balancing algorithms. The basic result will probably still hold, but differences in procedures will probably be smaller than those observed here to the extent that non-optimal item selection methods are used. Unique and complex item selection methods need to be investigated separately.

References

American College Testing Program (1993). *COMPASS User's Guide*. Iowa City, IA: ACT.

Bergstrom, B. A., Lunz, M. E. & Gershon, R. C. (1992). Altering the level of difficulty in computer adaptive testing. *Applied Measurement in Education, 5(2),* 137-149.

Divgi, D. R. (1991, September). *An analysis of CAT-ASVAB scores in the Marine Corps JPM data* (CRM 91-161). Alexandria, VA: Center for Naval Analysis.

Henly, S. J., Klebe, K. J. & McBride, J. R. (1989). Adaptive and conventional versions of the DAT: The first complete test battery. *Applied Psychological Measurement, 13(4),* 363-371.

Kingsbury, G. G. & Weiss, D. J. (1983). A comparison of IRT-based adaptive mastery testing and a sequential mastery testing procedure. In D. J. Weiss (Ed.) *New horizons in testing*. New York: Academic Press.

Lunz, M. E., Bergstrom, B. A. & Wright, B. D. (1992). The effect of review on student ability and test efficiency for computerized adaptive tests. *Applied Psychological Measurement, 16(1),* 33-40.

Mislevy, R. J. & Bock, R. D. (1983). *BILOG: Item and test scoring with binary logistic models*. Mooresville, IN: Scientific Software.

Owen, R. J. (1975). A Bayesian sequential procedure for quantal response in the context of adaptive mental testing. *Journal of the American Statistical Association, 70,* 351-356.

Parshall, C. G., Mittelholtz, D. & Miller, T. R. (1994, April). Response latency: an investigation into determinants of item-level timing. Paper presented at the national meeting of the National Council of Measurement in Education, New Orleans.

Reckase, M. D. (1983). A procedure for decision making using tailored testing. In D. J. Weiss (Ed.) *New horizons in testing*. New York: Academic Press.

Spray, J. A. & Reckase, M. D. (1993). Comparison of SPRT and sequential Bayes procedures for classifying examinees into two categories using an adaptive test. Unpublished manuscript.

Wald, A. (1947). *Sequential analysis*. New York: Wiley.

Ward, W. C. (1988). The College Board computerized placement tests: an application of computerized adaptive testing. *Machine-mediated Learning, 2*, 271-282.

Weiss, D. J. & Kingsbury, G. G. (1984). Application of computerized adaptive testing to educational problems. *Journal of Educational Measurement, 21*, 361-375.

Figure 1

Average Test Lengths Conditional on True θ
For the Sequential Bayes Procedure
Using Items Selected to Maximize Information
at the Decision Point, at True θ, and at Estimated θ

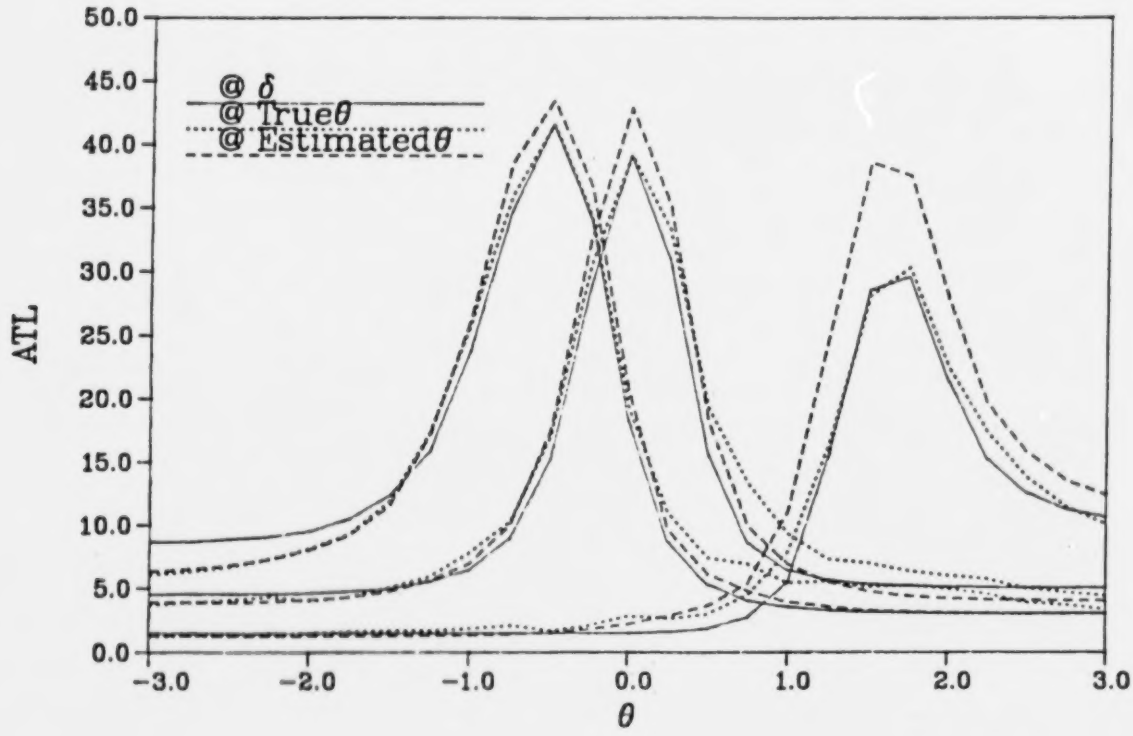

Figure 2

Average Test Lengths (ATL) Conditional on True θ
for the SPRT Procedure
Using Items Selected to Maximize Information
at the Decision Point or True θ